Information-Preserving Nonlinear Warping in Neural Networks

Abstract

Most widely used neural network activation functions introduce nonlinearity by destroying information. Examples include ReLU thresholding, sigmoid saturation, and max pooling. These operations partition or collapse regions of the input space and thereby prevent deep compositions of affine transforms from reducing to a single affine map.

However, information destruction is not mathematically necessary for nonlinear expressivity. A network can instead employ nonlinear but information-preserving vector-space warps between linear mixing stages. Such systems operate more like repeated geometric coordinate deformations than feature-thresholding cascades.

This note discusses the geometry of these systems and explains why invertible nonlinear warping does not collapse into a single equivalent affine transformation.

---

1. Linear Layers Alone Collapse

A purely linear network has the form

$$x \to A_1 x \to A_2 x \to \cdots \to A_n x$$

which simplifies to

$$x \to A_{\text{eff}} x$$

where

$$A_{\text{eff}} = A_n A_{n-1} \cdots A_1$$

Thus arbitrarily deep stacks of affine transforms remain globally affine.

The reason is that linear operators preserve:

- superposition,
- straight lines,
- global affine geometry.

No new geometric structure is introduced.

---

2. Standard Activations Prevent Collapse by Destroying Information

A conventional neural layer is

$$Y = \phi(Ax+b)$$

where ϕ is usually applied independently per coordinate.

Examples:

- ReLU,
- sigmoid,
- tanh,
- max pooling.

These operations are nonlinear because they violate additivity:

$$\phi(x+y) \neq \phi(x)+\phi(y)$$

but they are also generally many-to-one mappings.

For example:

$$\mathrm{ReLU}(x) = \max(0,x)$$

maps all negative inputs onto the same output region.

Thus standard networks obtain expressive power partly through:

- thresholding,
- region collapse,
- information deletion,
- piecewise partitioning of space.

---

3. Information Preservation Does Not Imply Affine Collapse

A common concern is that if nonlinearities preserve information, then the network may again reduce to a single affine operator.

This is false.

The key property required to avoid affine collapse is not information destruction but nonlinearity.

Consider an invertible nonlinear warp:

$$\Phi(x)$$

with

$$\Phi(x+y) \neq \Phi(x)+\Phi(y)$$

For example:

$$\Phi(x)=x+\alpha x^3$$

applied coordinatewise.

Now consider two linear stages separated by this warp:

$$y = B\Phi(Ax)$$

If this collapsed to a single affine map ($Cx+d$), then Φ would need to preserve superposition under arbitrary linear mixing, which nonlinear functions do not.

Equivalently:

$$B\Phi(A(x_1+x_2)) \neq B\Phi(Ax_1)+B\Phi(Ax_2)$$

in general.

Thus nonlinear invertible warping prevents affine collapse even while preserving information.

---

4. Geometric Interpretation

Linear layers define projection geometries:

- hyperplanes,
- subspace decompositions,
- coordinate mixing.

Standard activations modify these geometries through local thresholding.

Information-preserving nonlinear warps instead deform the coordinate system itself.

The network becomes:

1. linear projection,
2. geometric warp,
3. re-projection,
4. further warp,
5. etc.

This resembles repeated coordinate transport on a manifold rather than repeated threshold gating.

The essential computational resource is:

- curvature,
- non-uniform metric distortion,
- changing interpolation structure,

not information destruction.

---

5. Local vs Global Nonlinearities

Most activations are coordinatewise:

$$\Phi(x_1,\dots,x_n) = (\phi(x_1),\dots,\phi(x_n))$$

These produce axis-aligned distortions.

More general vector-to-vector nonlinearities

$$y = \Phi(x)$$

allow:

- norm-dependent warping,
- angle-sensitive transforms,
- subspace coupling,
- manifold-aware geometry,
- global coordinate adaptation.

This opens a larger design space in which nonlinear computation occurs through continuous geometric deformation rather than local thresholding.

---

6. Advantages of Information-Preserving Warps

Potential advantages include:

Reduced information loss

Representations may remain more invertible and distinguishable across depth.

Improved interpolation behavior

Intermediate representations may preserve smoother semantic structure.

Better gradient transport

Invertible or near-isometric mappings can reduce gradient degradation.

Global geometric reasoning

Vector-level nonlinearities can manipulate:

- norms,
- covariance,
- phase,
- spectral structure,
- distributed patterns.

Richer manifold transport

Computation becomes repeated reshaping of representation geometry.

---

7. Potential Difficulties

Several challenges arise:

Optimization complexity

Global nonlinearities create highly coupled gradients.

Reduced implicit regularization

Information destruction in ReLU-like systems acts as accidental denoising and sparsification.

Stability

Repeated nonlinear geometric warping may accumulate excessive curvature or chaotic sensitivity.

Computational cost

General vector nonlinearities may require $O(n^2)$ coupling operations.

---

8. Relation to Existing Directions

Several modern architectures partially explore related ideas:

- reversible networks,
- normalizing flows,
- orthogonal/unitary networks,
- scattering transforms,
- Hamiltonian neural systems,
- diffusion transport models.

However, most deep learning systems still fundamentally rely on information-destroying local nonlinearities.

The alternative perspective developed here is:

> Neural computation need not arise from repeated feature deletion and region collapse.
>
> It may instead arise from repeated information-preserving geometric warping of representation space.

This suggests a broader class of deep architectures based on nonlinear coordinate transport rather than threshold partitioning.